IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1891∼1908
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1891-1908

TunDC: a public benchmark dataset for sentiment analysis and language modeling in the Tunisian dialect

Ahmed Khalil Boulahia[1], Mourad Mars[2]
[1]Tunis Dauphine University, Tunis, Tunisia
[2]Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Mecca, Saudi Arabia

Article Info

Article history:

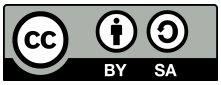
Received May 8, 2025
Revised Jan 29, 2026
Accepted Feb 6, 2026

Keywords:

Arabic dataset
Artificial intelligence
Fine-tuning
Large language model
Low-resource language
Sentiment analysis
Tunisian dialect

ABSTRACT

The development of natural language processing (NLP) applications has increasingly focused on dialectal variations of languages. The Tunisian dialect (TD), a widely spoken variant of Arabic, poses unique linguistic challenges due to its lack of standardized writing conventions and influences from multiple languages, including French, Italian, Turkish, and Berber. In this work, we introduce TunDC, a dataset of 20,044 labeled comments designed to advance NLP research on the TD. The dataset covers diverse linguistic forms (Arabic, Latin, and mixed scripts), and each comment was manually annotated for positive or negative sentiment by native speakers, achieving high inter-annotator agreement. To evaluate its effectiveness, we fine-tuned various models on TunDC. The bert-base-arabic-TunDC-mixed model achieved an accuracy of 0.84 and a macro-averaged F1-score of 0.83, demonstrating strong generalization across sentiment categories and writing systems. A stratified data-splitting strategy considering both sentiment and script type further improved accuracy by approximately 8% compared to standard splits. As a publicly available resource, TunDC contributes to the computational linguistics community, fostering advancements in language modeling and applications tailored to the TD.



Corresponding Author:

Mourad Mars
Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University
Makkah 24382, Saudi Arabia
Email: msmars@uqu.edu.sa

1. INTRODUCTION

Recent years have witnessed remarkable advancements in large language models (LLMs) and generative AI, leading to significant breakthroughs in various natural language processing (NLP) tasks. From machine translation and text summarization to writing different kinds of creative content, these models have demonstrated exceptional capabilities in understanding and generating human language. However, a major obstacle to broader adoption and linguistic inclusivity is the persistent scarcity of annotated data, especially for low-resource languages and dialects. This limitation often leads to significant performance gaps, undermining the linguistic richness of these communities and restricting their access to NLP technologies.

One such case is Tunisian Arabic (TA), an Arabic dialect spoken by over 12 million individuals in Tunisia, which offers a captivating and underexplored aspect. Unlike standard Arabic, it possesses unique linguistic characteristics shaped by historical and cultural influences. Its vocabulary draws heavily from Arabic but also incorporates a wide range of loanwords from Berber, Turkish, French, English, and Italian, enriching

its expressive power and reflecting its vibrant sociolinguistic landscape (Table 1). Yet, the lack of large labeled datasets tailored for sentiment analysis in the Tunisian dialect (TD) poses a barrier to deeper exploration and hinders the development of robust NLP applications.

When it comes to writing in the TD, people use different writing systems: Arabic script (abjad), Latin script, and even numbers to represent some characters. Most of the time, the TD doesn't have any well-defined structure and does not conform to any conventions or orthographic rules. These characteristics of the TD present an additional challenge in effectively processing it using NLP techniques. Table 2 present some of most common expressions that represent translation of English sentence "when are you going to the doctor?".

Table 1. Examples of loanwords in the TD

Tunisian dialect	Arabic	Origin	English translation
بانكا	بنك	Italian "banca"	Bank
برطمان	شقة	French "appartement"	Apartment
لاباس	جيد	Berber "labes"	Fine
تليفون	هاتف	English "talifoun"	Telephone

Table 2. Different TD expressions that represent a translation of "when are you going to the doctor?"

Reference sentence	When are you going to the doctor?
Sentence 1	وقتاش تمشي للطبيب؟
Sentence 2	Wa9tash timshi litbib?
Sentence 3	Waktech temchi lel doctour?
Sentence 4	Ana wa9t machi 3and tbib?

The processing of the TD represents a challenging task, mainly due to its ambiguous and complex structure, not only for machines but sometimes for Tunisians themselves. People can write in both directions (right to left and left to right) using the Arabic alphabet (abjad) and the Latin alphabet. On many occasions, they use both simultaneously. It also varies according to a person's age, origins, and culture. The majority of Arabic users on social media platforms use dialects to express themselves; most of these dialects can be described as unstructured, non-grammatical slang Arabic. This non-uniformity in dialects makes it more difficult for machine learning (ML) algorithms and LLMs to be able to perform tasks such as sentiment analysis. Hence, there is an increasing need for larger datasets to improve the performance of these models.

This paper aims to tackle this challenge by introducing TunDC, a new benchmark corpus specifically designed for sentiment analysis in TA. We leverage the power of social media, collecting and annotating more than 20K comments with sentiment labels (positive or negative) provided by native speakers. TunDC is intended not only as a training resource but also as a standardized benchmark for evaluating and comparing NLP systems on TD sentiment analysis, with potential use in future shared tasks and competitions. Moreover, the dataset's scale and script diversity make it suitable for pre-training dialect-specific language models and for enabling effective transfer learning through fine-tuning of multilingual or Arabic-centric transformers such as Arabic bidirectional encoder representations from transformers (AraBERT), multidialectal Arabic BERT (MARBERT), or cross-lingual language model-robustly optimized BERT pre-training approach (XLM-R). By addressing the data scarcity issue, TunDC aims to empower future research and development of sentiment analysis solutions tailored to the unique linguistic characteristics of TA [1], [2].

This paper makes three key contributions. First, it provides a survey of available resources for the TD. Second, it introduces TunDC, a novel, publicly available benchmark dataset for sentiment analysis in TA, designed to support both model training and standardized evaluation. Third, it presents the training, evaluation, and public release on Huggingface of multiple pre-trained LLMs fine-tuned via transfer learning, including Camembert-BERT (CamemBERT) (AhmedBou/camembert-TunDC), bert-base-uncased (AhmedBou/camembert-TunDC), Modernized BERT (ModernBERT) (AhmedBou/ModernBERT-TunDC), and bert-base-arabic (AhmedBou/bert-base-arabic-TunDC-mixed), on the TunDC dataset.

The rest of the paper is organized as follows: section 2 provides an overview of related work in sentiment analysis and datasets for Arabic dialects. Section 3 details the dataset creation process, covering data collection, preprocessing, annotation, statistical analysis, and evaluation. Section 4 describes the experimental setup, presents the results, and discusses the findings. Section 5 discusses ethical considerations, including bias and fairness in Tunisian dialect NLP. Finally, section 6 concludes the paper and outlines potential directions for future research.

2. RELATED WORKS

Sentiment analysis in the TD presents a unique challenge due to the dialect's distinct vocabulary, morphology, and syntax compared to modern standard Arabic (MSA) [3], [4]. Several research efforts have focused on addressing this challenge, employing diverse approaches and datasets [5]–[11]. This section provides a comprehensive overview of existing work, with a particular focus on recent advances in multilingual models, dialect adaptation, and the specific challenges and opportunities within TA NLP.

2.1. Advances in multilingual transformers and dialect adaptation

The advent of the Transformer architecture has revolutionized NLP, with multilingual large language models (MLLMs) like XLM-R and the more recent generative models demonstrating impressive cross-lingual capabilities. Recent work has focused on adapting these powerful models to low-resource dialects. A key challenge is the dialect adaptation of these MLLMs. While models pre-trained on massive amounts of data, including some Arabic content, show a baseline performance, they often struggle with the nuances of specific dialects like TD. Studies such as those presented at Arabic natural language processing 2025 (ArabicNLP 2025) and empirical methods in natural language processing 2025 (EMNLP 2025) have shown that MLLMs exhibit a significant "Arabic gap" in handling dialectal variations, code-mixing, and script-switching [12].

Model merging and continual pre-training have emerged as effective strategies for dialect adaptation, moving beyond conventional fine-tuning. For instance, research in 2025 explored model merging to adapt multilingual models for code-mixed tasks, a highly relevant challenge for TD which frequently mixes Arabic script, Latin script (Arabizi), and French loanwords [13]. Furthermore, the evaluation of generative LLMs in zero-shot and few-shot settings for Arabic tasks, including sense disambiguation and translation, highlights the growing trend of leveraging the inherent knowledge within these models to overcome data scarcity in dialectal NLP [14], [15].

2.2. Zero- and few-shot learning for low-resource dialects

The scarcity of large, high-quality annotated datasets for dialects necessitates the exploration of zero-shot (ZSL) and few-shot learning (FSL) techniques. These methods are crucial for advancing NLP in low-resource settings, including TA. Recent surveys (2024-2025) on Arabic dialect processing emphasize the shift towards ZSL and FSL, often utilizing the power of LLMs. In the context of sentiment analysis, ZSL and FSL allow models to generalize from instructions or a handful of examples, bypassing the need for extensive, costly manual annotation [16]. For TD specifically, researchers are benchmarking LLMs' performance, finding that while they struggle initially, prompting techniques and FSL can significantly improve their ability to recognize and adhere to the dialect's unique linguistic structure [17]. This approach is particularly promising for tasks like sentiment analysis where the underlying sentiment concept is universal, but the dialectal expression is unique.

2.3. Technical gaps and linguistic challenges in Tunisian dialect natural language processing

The TD presents a unique set of linguistic and technical challenges that impede the development of robust NLP systems. The most significant linguistic hurdle is the script-switching phenomenon, where users fluidly switch between Arabic and Latin characters, often within the same sentence, to represent TD phonemes. This necessitates resources like Tunisian Arabish corpus (TArC) [18] and the newly introduced LinTO datasets [19], which focus on transliteration and linguistic annotation to bridge the gap between the written forms. The technical gap is primarily the lack of a large-scale, multi-domain, and multi-script benchmark dataset, which TunDC aims to address. As shown in Table 3, extreme code-switching and script-mixing complicate model design, while the lack of orthographic standardization increases lexical variability. On the technical side, data scarcity and domain adaptation challenges further limit the performance of NLP systems, motivating the development of the TunDC dataset.

2.4. Arabic regional dialects

Unlike English, Arabic language presents additional challenges due to its multiple dialects, the limited availability of large corpora, and the absence of vocalization. Therefore, creating high-quality datasets and developing NLP tools capable of accurately processing dialectal Arabic is crucial. Significant efforts have been made to construct datasets for specific dialects, with the Egyptian (EGY) [20]–[22] and Levantine (LEV) dialects being the most extensively studied [23], [24]. More recently, research has expanded to include the Palestinian (PAL) [25], Khaliji [26], [27], Syro-Palestinian [22], Gulf (GLF) [28], Mesopotamian (Iraqi) [22],

and Maghrebi (MGR) [29] dialects [30], [31]. However, the TD remains underexplored, with limited linguistic resources and NLP tools available.

Table 3. Technical gaps and linguistic challenges in TA NLP

Challenge type	Specific challenge in TD	Impact on NLP development
Linguistic	Extreme code-switching/mixing: frequent, unstandardized mixing of Arabic script, Latin script (Arabizi), and French/Italian/Turkish loanwords.	Requires models to handle multiple orthographies and languages simultaneously, increasing model complexity and data requirements.
Linguistic	Lack of orthographic standardization: no fixed rules for writing TA, leading to high lexical variability (e.g., multiple ways to write the same word).	Hinders the effectiveness of traditional tokenization, stemming, and lexicon-based methods.
Technical	Data scarcity and fragmentation: limited availability of large, publicly accessible, and high-quality annotated corpora. Existing datasets are often small and task-specific.	Prevents the effective pre-training of dedicated, high-performing TD language models.
Technical	Domain adaptation: models trained on one domain (e.g., political tweets) perform poorly on others (e.g., e-commerce comments).	Requires continuous adaptation strategies and diverse datasets like TunDC to ensure generalizability.

2.5. Tunisian dialect datasets

The first interest in TD sentiment analysis was in 2016, Sayadi *et al.* [32] presented in their paper, a sentiment analysis study on the first labeled and publicly available dataset called Tunisian election corpus (TEC). This dataset is composed of 5,514 tweets collected during the Tunisian elections period of 2014. 3,760 of them are in MSA, and 1,754 are in the TD. Many ML approaches were presented, and a comparative study was conducted. The presented results showed that support vector machines (SVM) achieved a higher accuracy of 71.09% than the other methods used.

The Tunisian Arabic corpus (TAC) [33] consists of 800 tweets covering various topics, including media, telecommunications, and politics. This dataset was gathered by Karmani [33] and labeled with sentiment categories: positive, negative, and neutral. In 2017, a dataset called the Tunisian sentiment analysis corpus (TSAC) was presented and made available publicly for the NLP Tunisian community [34]. The dataset was obtained from Facebook comments about popular TV shows, and it is written only with Arabic letters. The authors reported the first application of deep learning in sentiment analysis on the TD, where they used multi-layer perceptron (MLP), which produced a lower error rate than SVM and naïve-Bayes and reached 78% accuracy.

Tunisian Arabizi (TUNIZI) [35] contains 9,210 comments gathered from the YouTube platform and labeled positive and negative. Many topics were covered in this dataset, such as sports, politics, comedy, and TV shows. Both classes are similarly represented in this dataset, with 47% positive comments and 53% negative comments. Masmoudi *et al.* [36] introduced a manually annotated dataset for sentiment analysis of the TD, composed of comments collected from official Facebook pages of Tunisian supermarkets. The dataset was labeled based on five sentiment categories (very positive, positive, neutral, negative, and very negative) and twenty aspect-based categories. To analyze sentiment, the authors experimented with three deep learning models: convolutional neural networks (CNN), long short-term memory (LSTM), and bidirectional long short-term memory (Bi-LSTM). Their evaluation showed that CNN and Bi-LSTM achieved the best classification performance, demonstrating the effectiveness of deep learning in processing TD text.

Gugliotta and Dinarelli [18] introduced TArC, a publicly available dataset designed for processing TA written in Arabizi. The corpus was developed alongside an NLP tool that provides various levels of linguistic annotation, including word classification, transliteration, tokenization, part-of-speech tagging (POS-tagging), and lemmatization. The authors outlined their computational and linguistic methodologies, discussing strategies to enhance annotation accuracy. Their experiments demonstrated the effectiveness of these resources for both computational applications and linguistic research.

Mulki *et al.* [37] investigated sentiment analysis of the TD using both supervised and lexicon-based models. They evaluated preprocessing techniques such as stemming, emoji recognition, and negation detection on three datasets of varying sizes. Their results showed that these preprocessing steps significantly improved sentiment classification performance, with named entity tagging further enhancing lexicon-based models and benefiting supervised models on smaller datasets.

To summarize, Table 4 provides an overview of the datasets collected for the TD sentiment analysis task. The existing datasets are either small, limited in script coverage, or focused on specific tasks, such as transliteration. TunDC distinguishes itself by offering a large, publicly available, and script-diverse corpus (Arabic, Latin, and mixed scripts) with high-quality, manually verified sentiment annotations, positioning it as a robust benchmark for multi-script and multi-domain sentiment analysis.

Table 4. Summary of available datasets for TD sentiment analysis

Study	Dataset	Size	Source	Labels
Sayadi *et al.* [32]	TEC [32]	5,514 tweets	Twitter	Positive, Negative
Karmani [33]	TAC [33]	800 tweets	Facebook	Positive, Negative, Neutral
Medhaffar *et al.* [34]	TSAC [34]	17k tweets	Facebook	Positive, Negative
Fourati *et al.* [35]	TUNIZI [35]	9,210 YouTube comments	YouTube	Positive, Negative
Masmoudi *et al.* [36]	Comments about	17k Arabic script posts	Facebook	Very Positive
				Positive, Neutral
	Tunisian supermarkets [36]	27k Arabizi script posts		Negative, Very Negative
Gugliotta and Dinarelli [18]	TArC [18]	11,291 comments	Facebook	Positive, Negative, Neutral

3. TUNISIAN DIALECT CORPUS

TunDC was developed through a three-step process consisting of data gathering, content filtering and preprocessing, and one-stage annotation. This dataset is designed to be diverse and well-structured. Providing a valuable resource for training and evaluating deep learning models for sentiment analysis in the TD.

3.1. Data gathering, pre-processing, and labelling

A further inspection of the TD typesetting, commonly used on social media, shows that there are many factors affecting its structure, such as the user's age, sex, region, and interests. The goal is to build a large dataset that contains the majority of this used vocabulary. We have gathered approximately 24K comments through data scraping from over 300 social media posts and videos across various Tunisian Facebook pages and YouTube channels, with the most recent scraping conducted in January 2025. These comments cover a broad array of topics, including music, politics, sports, news, and TV shows. To ensure vocabulary diversity, we limited the maximum number of comments scraped per post to 200. Additionally, we intentionally included longer comments in the dataset to better process and understand extended contextual entries. The longest comment in the TunDC dataset spans 873 tokens, while the average comment length is 42 tokens.

We filtered out the non-TD comments and comments fully written in MSA using fastext scoring [38], [39]. Then, other filters were performed on the collected dataset to exclude inappropriate comments (harmful and offensive) and to involve only the comments from Tunisian people that were written in TD. Additionally, comments consisting of only one word were excluded from the dataset to maintain the quality and relevance of the content. This step ensures that the remaining comments contain sufficient context and information, allowing for more accurate sentiment analysis. The last step in filtering was to perform deduplication [40] by removing near-duplicate examples and long repetitive sub-strings to improve the quality of the dataset allowing for more accurate evaluation and better model performance [41]–[43]. The complete step-by-step filtering, script classification, and sentiment labeling process applied to the TunDC dataset is illustrated in Figure 1.

After the gathering and filtering phases, the data preprocessing starts by removing all kinds of links and special characters. Recognizing that emojis carry significant semantic meaning and are prevalent in real-world text, we intentionally preserved them during preprocessing, rather than replacing them with decoded formats (as might be done with packages like the emoji package [44]). This approach aims to improve the model's understanding of nuanced expressions and enhance its generalization capabilities on authentic data.

To ensure the accuracy and reliability of the labels, a careful labeling process was implemented. Initially, a representative sample of approximately 20K comments was carefully selected, ensuring the inclusion of diverse linguistic styles, topics, and sentiment expressions. To establish a consistent labeling framework, a comprehensive description of labels, detailed guidelines, and a list of key instructions were provided. In addition, annotated comment examples were presented to the annotators to foster a clear understanding of the labeling criteria. The primary labeling task involved determining whether each comment conveyed a positive or negative sentiment.

Each comment underwent a three-stage annotation process that involved three native Arabic and Tunisian speakers volunteering as annotators. In the first stage, each annotator independently assigned a sentiment label, either positive or negative, to the comment. In the second stage, the labels assigned by the three annotators were compared. If three or two annotators agreed on the sentiment label, that label was considered the final annotation for the comment. This majority vote approach ensured that the final sentiment labels were consistent and reflected the consensus of the annotators. However, if one of the three annotators was uncertain about the labeling of the comment, the comment was excluded from the dataset. This exclusion step ensured that only comments with clear and consistent sentiment labels were retained, further enhancing the quality of the dataset. The agreement between our three annotators, measured using the Fleiss Kappa measurement, was almost perfect (kappa =0.97) [45], [46].

The ambiguity in labeling often arose from comments containing sarcasm, rhetorical questions, quotes from others, or mixed sentiments within a single statement. Such nuances made it challenging for annotators to assign a clear positive or negative label consistently. Translating these comments into English is especially challenging, as much of their meaning, tone, and cultural context would be lost. Table 5 provides several examples illustrating these cases. We constructed TunDC, a sentiment analysis corpus for TA, comprising 9,088 positive and 10,956 negative examples, ensuring a good distribution for robust ML and language modeling applications.

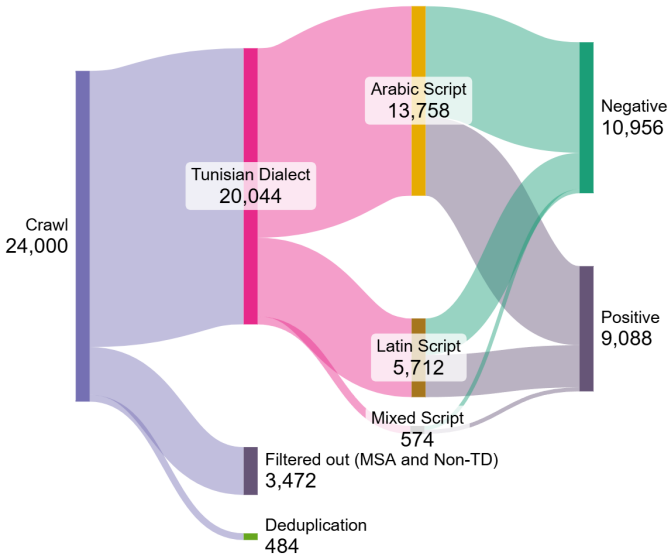


Figure 1. Step-by-step filtering, script classification, and sentiment labeling process applied to TunDC dataset

Table 5. Examples of ambiguous and borderline sentiment cases and their assigned labels

Ambiguous examples	Challenge	Assigned label
يرحم والديه كان جو منو برشا في تونس رانا لباس (May God bless his parents if Tunisia had more people like him, things would be much better for us.)	Mixed sentiment, colloquial nuances	Positive
مستقبل البلاد من مستقبل شبابها . أنقذوا شباب تونس فهي مسؤولية الدولة (The future of the country depends on the future of its youth. Save Tunisia's youth, it is the responsibility of the state.)	General statement with implied sentiment	Positive
Cest bien bech hata rbo3 rajil maytjara wymed yedo 3la marto ydhrbha (It's great, so that even a quarter of a man won't dare and raise his hand to beat his wife.)	Sarcasm, indirect expression	Negative
التاريخ باهي كي نعرفوه اما موش على حساب الحاضر و المستقبل (History is important to know, but not at the expense of the present and future.)	Statement with contrasting ideas	Positive
Brabbi ya mosaique iktbou 7aja s7i7a. Martou hia awal 3amla fi l Amazon wa wa9fit el charika 3la sa9iha m3ah. Ya3ni chrika 50/50 kan mouch akthar (Please, Mosaique, write something accurate. His wife was the first employee at Amazon and helped build the company with him. It's a 50/50 company, if not more.)	Long, detailed comment with mixed facts and opinions	Negative

3.2. TunDC dataset description

TunDC's primary purpose is to facilitate sentiment analysis tasks. To achieve this goal, we crafted a near-balanced dataset in terms of the proportion of writing systems (Arabic and Latin) within each sentiment label class. This representation ensures that the dataset accurately reflects the distribution of writing systems in real-world TD usage, enabling sentiment analysis models to effectively capture sentiment patterns across both writing conventions. The labels are represented as follows: 45.3% are positive labels and 54.6% are negative labels. The comments written in Arabic alphabet represents around 68.6% of the total data, whereas the comments written in Latin alphabet only or mixed represent the remaining 31.4%. To the best of our knowledge, this is the largest public TD dataset with over 20K manually labeled comments. Table 6 presents the exact distribution of comments in the dataset, categorizing them by sentiment (positive and negative) and writing system (Arabic, Latin, and mixed).

Table 6. Distribution of comments by sentiment and writing system in TunDC

TunDC	# Comments	# Arabic comments	# Latin comments	# Mixed comments
Positive comments	9,088	5,735	3,061	292
Negative comments	10,956	8,023	2,651	282

Figure 2 illustrates the distribution of each writing system (Arabic and Latin) per label (positive or negative). It displays the distribution of comments written in TD across positive and negative sentiment labels. The x-axis represents the sentiment label (positive and negative), while the y-axis denotes the number of comments. The results indicate that approximately 55% of the Arabic comments express a negative sentiment, whereas 45% convey a positive sentiment. Figure 2 also presents a donut chart of the distribution of comments written in Latin script across positive and negative sentiment labels. The data reveal that around 60% of the Latin comments are positive, while 40% are negative. This pattern aligns with the sentiment distribution observed in Arabic-script comments, indicating a similar sentiment trend across both writing systems. These figures clearly illustrate that the distribution of writing systems is almost balanced across both sentiment labels. This balanced representation is crucial for ensuring that the sentiment analysis models trained on TunDC can accurately capture sentiment patterns across both writing conventions and sentiment classes. The following analysis examines the word and token distribution within the dataset, offering insights into its composition and suitability for different NLP tasks.

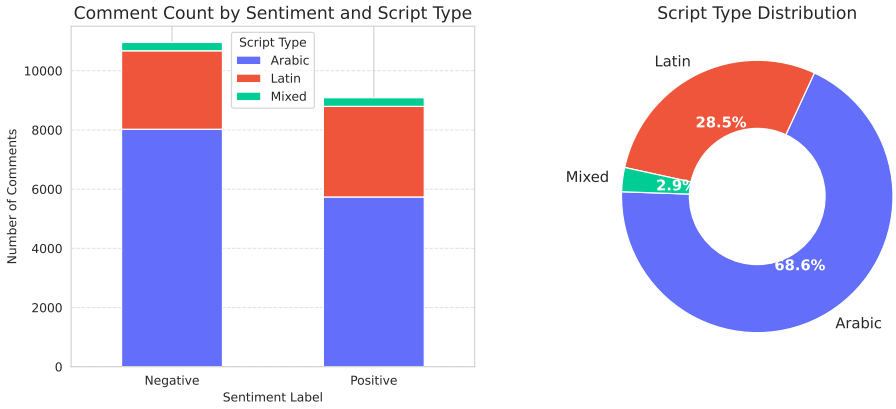


Figure 2. Distribution of positive and negative comments across writing systems (Arabic, Latin, and Mixed) in the TunDC dataset, along with the overall proportion of each script type

The word count distribution (Figure 3) highlights a dominant range between 2 and 10 words, showing that short entries make up the majority of the dataset. The alignment between token and word counts follows typical patterns, with slight variations due to tokenization complexities, such as special characters or encoding rules. This distribution implies a dataset that is well-suited for classification or lightweight NLP tasks rather than generation-heavy applications. Expanding the dataset to include longer texts may help improve versatility if required for more comprehensive NLP solutions.

Similarly, the token distribution plot (Figure 4) reveals a skewed pattern, with the majority of entries containing fewer than 50 tokens. This indicates that most texts in the dataset are concise, with a sharp decline in frequency as token counts increase. The cl100k_base tokenizer appears to efficiently compress text, as expected, with relatively short token sequences dominating the data. The steep drop-off after 50 tokens suggests that long-form content is rare, possibly reflecting a dataset built around brief comments or text snippets.

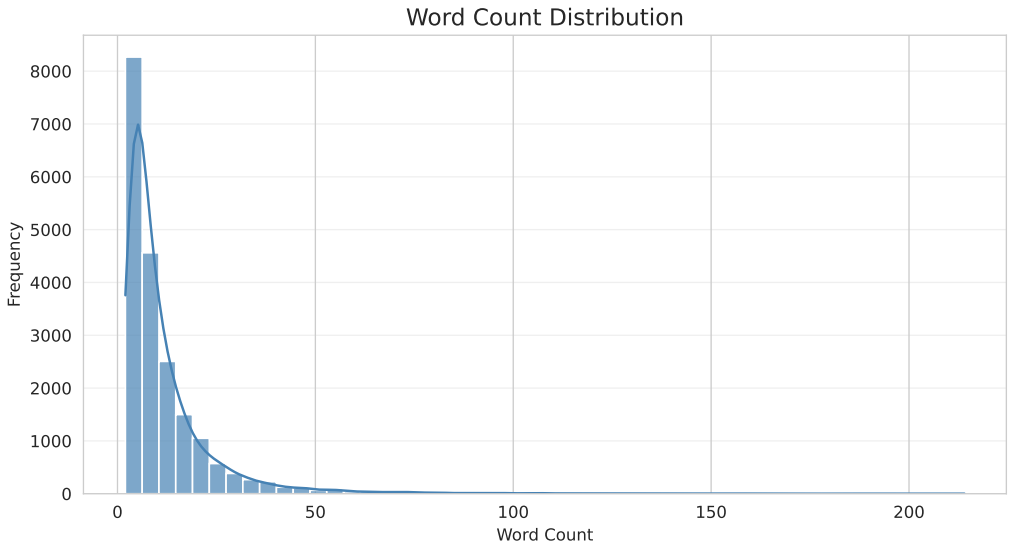


Figure 3. The word count distribution in TunDC

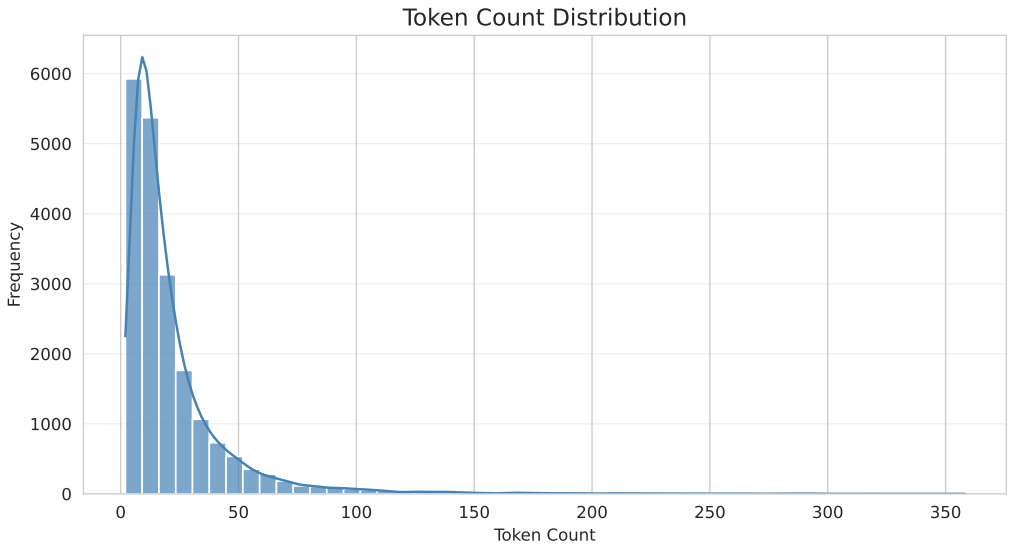


Figure 4. The token distribution in TunDC

In addition to the overall word and token distributions, we also analyzed the lexical diversity of the TunDC dataset. Specifically, TunDC contains 71,372 unique words, highlighting the broad range of expressions used in TD content. Among these, 47,889 unique words are in Arabic script, while 21,306 unique words are written in Latin script. This considerable presence of Latin-scripted terms reflects the multilingual and code-switching nature of TA, which frequently blends elements from French, English, and other languages into everyday digital communication. Such diversity underscores the linguistic richness of the dataset and reinforces the importance of tailoring NLP models to effectively handle both scripts in low-resource dialects.

Table 7 presents a selection of randomly chosen comments from the TunDC dataset. These comments represent a diverse range of sentiment labels and writing systems (Arabic, Latin, and mixed), providing insight into the linguistic and contextual variations within the dataset. By showcasing real examples, this table highlights the complexity of sentiment analysis in the TD and the challenges associated with processing dialectal Arabic text. Table 8 provides a comparison of TunDC's key features and characteristics with other publicly available TD datasets.

Table 7. Sample of randomly selected comments from TunDC dataset

Script type	Comment	Translation	Label
Arabic	مليون مزال تعمل حاجة اصلا.. هاذي منحة فقر موش شهرية	A million barely gets you anything... This is a poverty handout, not a proper salary.	Negative
Arabic	مسلسل روعة..ممثلين ممتازين ..سيناريو هبال برافوووووا برشا ابداع	What a fantastic series! Excellent actors... The script is absolutely crazy good. Bravo! So much creativity!	Positive
Latin	*9adeh 5alsouk ya ta7foun bech tahki Hal kelmtin li kif wejhek*	How much did they pay you, cutie, just to say those two worthless words?	Negative
Latin	*wallahi nhebek barcha wo n9adrek ena*	I swear to God, I love you very much and I respect/appreciate you.	Positive
Mixed	*Fils à maman* ....وما اكثرهم	Mama's boy, and there are so many of them...	Negative
Mixed	Bravo قرار في محله	That's a sound decision. Bravo	Positive

Table 8. Comparing TunDC to other publicly available sentiment datasets

Feature	Dataset size	Label classes	Data source	#Pos	#Neg
TunDC	20,044	2	Facebook and Youtube	9,088	10,956
TEC [32]	5,514	2	X (Twitter)	-	-
TSAC [34]	17K	2	Facebook	8,845	8,215
TUNIZI [35]	9,210	2	Youtube	4,372	4,838

4. RESULTS AND DISCUSSION

This section presents the experimental setup and results to evaluate the quality and utility of the TunDC dataset for sentiment classification in TD NLP. We assess the performance of a transformer-based model, fine-tuned on TunDC, to determine its effectiveness in capturing dialectal nuances. Key evaluation metrics, including accuracy, precision, recall, and F1-score, are used to analyze model performance. The following subsections outline the dataset preprocessing, model setup, evaluation criteria, and training-testing procedures.

4.1. Experimental setup

4.1.1. Dataset preparation

The TunDC dataset was divided into an 85% training set and a 15% validation set. We employed a stratified splitting strategy based not only on sentiment classes (positive and negative) but also explicitly on writing systems (Arabic, Latin, and mixed). This multi-criteria stratification approach proved effective, contributing to an approximate 8% increase in accuracy compared to simpler splits. Subsequent preprocessing focused on data cleaning and normalization. Specific cleaning steps included: filtering out comments consisting of only one word to ensure meaningful content; removing all URLs and user mentions for anonymization and noise reduction; and retaining emojis and hashtags if they appeared alongside text, acknowledging their significant semantic and emotional value. These preprocessing steps were essential for improving the model's ability to generalize across diverse linguistic variations in the TD.

4.1.2. Baseline models

Classical approaches: to contextualize the performance of our transformer-based solution, we established a set of classical ML baselines spanning distinct architectural paradigms: a SVM, an ensemble gradient-boosted tree model (extreme gradient boosting (XGBoost)), and a lightweight recurrent neural network (a single Bi-LSTM layer with 16 units followed by a dense layer of 8 units). All models were trained using default hyperparameters and evaluated on the same stratified 15% test split of the TunDC dataset described above, ensuring a fair comparison. No architecture-specific tuning was performed, allowing us to assess out-of-the-box performance under identical data conditions. Table 9 summarizes their results on the test set in terms of accuracy, F1-score, and area under the curve (AUC).

Table 9. Performance of classical baseline models on the TunDC test set

Model	Accuracy	F1-score	AUC
SVM	0.77	0.77	0.85
XGBoost	0.76	0.75	0.84
LSTM (Bi-LSTM-16 + Dense-8)	0.78	0.76	0.85

Transformer-based approaches: we further evaluated four pre-trained ArabBERT-style Transformer models, each trained on different combinations of MSA and dialectal content, to assess their out-of-the-box suitability for TD sentiment analysis. The selected models include: bert-base-arabic-camelbert-da (dialect-focused), the second model is the UBC-NLP/MARBERTv2 (trained on dialect-heavy social media text), CAMeL-Lab/bert-base-arabic-camelbert-mix (mixed MSA and dialect), and the last model is the asafaya/bert-base-arabic (primarily MSA with some dialectal coverage). All models were fine-tuned identically using the same training protocol (learning rate $3e{-}5$, batch size 8, 4 epochs) and evaluated on the same stratified test split of TunDC. As shown in Table 10, all four models significantly outperformed classical baselines, with test F1-scores ranging from 0.826 to 0.837. Notably, both camelbert-mix and asafaya/bert-base-arabic achieved the highest evaluation accuracy and F1-score (0.837). Given its strong performance, broader linguistic coverage, and established use as a foundation for Arabic NLP tasks, we selected asafaya/bert-base-arabic as the base architecture for our subsequent hyperparameter tuning, error analysis, and final model development—leading to the TunDC-mixed model described in the next section.

Table 10. Performance of transformer-based baseline models on the TunDC test set

Model	Test accuracy	Test F1-score
camelbert-da	0.826	0.826
MARBERTv2	0.836	0.835
camelbert-mix	0.837	0.837
asafaya/bert-base-arabic	0.837	0.837

4.1.3. Evaluated model

For this study, we evaluated TunDC-mixed [47], our fine-tuned transformer-based model built upon asafaya/bert-base-arabic [48], [49]. To optimize performance on the TunDC dataset, we conducted systematic hyperparameter tuning using Ray Tune from the Ray library. Specifically, we employed random search over a predefined hyperparameter space, exploring the following dimensions:

- learning_rate: log-uniform sampling between $1e{-}5$ and $5e{-}5$.
- per_device_train_batch_size: categorical selection from {4, 8, 16}.
- weight_decay: uniform sampling in $[0.0, 0.1]$.
- num_train_epochs: integer values from 3 to 6.
- lr_scheduler_type: categorical choice between linear and cosine.

The search yielded the following optimal configuration: learning rate $3e{-}5$, batch size 8, weight decay 0.017, 4 training epochs, and a linear learning rate scheduler. Using these hyperparameters, the final model—comprising 111 million parameters—was trained with the AdamW optimizer. Its strong performance underscores its capacity to capture the linguistic nuances and contextual variability inherent in low-resource TA dialects, making it well-suited for sentiment classification in this domain.

4.1.4. Evaluation metrics

The model's performance was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score. Accuracy measured the overall correctness of predictions, while precision evaluated the proportion of correctly predicted positive (or negative) labels. Recall measured the model's ability to capture all relevant instances of a sentiment class. The F1-score, as the harmonic mean of precision and recall, ensured a balanced evaluation, particularly useful when handling variations in sentiment representation.

Additionally, macro-averaged F1-score was reported to treat each sentiment class equally, regardless of class size. The weighted F1-score was computed to account for class distribution, providing a more representative measure of overall model performance. These evaluation metrics offered a robust assessment of bert-base-arabic-TunDC-mixed in processing TD text.

4.1.5. Training and testing procedures

The training and testing process followed a structured pipeline to ensure robust model evaluation. PyTorch 2.5.1+cu124 and the Transformers 4.48.3 library were used for training, with tokenization handled by Tokenizers 0.21.0 to maintain consistency across different writing systems. Hyperparameter tuning was conducted on the validation set to optimize learning rates, batch sizes, and regularization parameters, ensuring stable training. The model was trained using AdamW optimization, with betas set to (0.9, 0.999) and an epsilon value of 1e-08. A linear learning rate scheduler was implemented to adjust the learning rate dynamically throughout training. To prevent overfitting, early stopping was used, monitoring validation loss over multiple epochs. The final model evaluations were performed on the test set, providing an objective assessment of its generalization ability in sentiment classification for the TD.

4.2. Results

The final TunDC-mixed model was evaluated on both the training and test splits of the TunDC dataset. On the test set, the model achieved an accuracy of 0.83 and an AUC of 0.9092, while on the training set, it reached an accuracy of 0.84 and an AUC of 0.9096—indicating strong discriminative capability with minimal overfitting. These results confirm the model's robustness in distinguishing between positive and negative sentiment in Tunisian dialectal text.

A per-class analysis reveals consistent behavior across splits. For the negative class (class 0), the model exhibits high precision (0.87 on both train and test) but slightly lower recall on the test set (0.81 vs. 0.82 on train), suggesting a small increase in false negatives during evaluation. Conversely, for the positive class (class 1), recall remains strong in both splits (0.85 train, 0.86 test), while precision drops modestly from 0.80 (train) to 0.79 (test), reflecting a slight rise in false positives. Overall, the F1-scores remain balanced—0.84 (train) and 0.83 (test) macro averages—demonstrating stable performance across sentiment categories. To provide a comprehensive view, we report detailed metrics in two tables. Table 11 presents class-wise precision, recall, F1-score, and support for both training and test sets. Table 12 summarizes the overall performance, including accuracy, macro and weighted averages, and AUC.

Table 11. Class performance metrics on training and test sets

	Training set			Test set		
Class	Precision	Recall	F1-score	Precision	Recall	F1-score
Negative (0)	0.87	0.82	0.85	0.87	0.81	0.84
Positive (1)	0.80	0.85	0.83	0.79	0.86	0.82

Table 12. Overall performance metrics including AUC

Metric	Training set	Test set
Accuracy	0.84	0.83
Macro Avg F1	0.84	0.83
Weighted Avg F1	0.84	0.83
AUC	0.9096	0.9092

To further illustrate the model's behavior, we present visual diagnostics in the form of confusion matrices and receiver operating characteristic (ROC) curves. Figure 5 displays the confusion matrices for both training and test sets, offering insight into the types and frequencies of classification errors. The model exhibits a slight tendency to misclassify positive instances as negative on the test set—a pattern consistent with the lower precision observed for the positive class. Complementing this, Figure 6 shows the ROC curves for both splits, with near-identical curves indicating consistent discriminative performance across training and evaluation. The high area under curve (AUC) values (0.9096 train, 0.9092 test) confirm the model's strong ability to distinguish between sentiment classes across decision thresholds.

4.3. Error analysis

To better understand the limitations of TunDC-mixed, we conducted a detailed error analysis exclusively on the test set. Overall, misclassification rates varied significantly across writing systems: 13.4% of Arabic-script comments, 25.9% of Latin-script comments, and 17.4% of mixed-script comments were incorrectly labeled. This disparity suggests that script choice—particularly Latin transcription—poses a unique challenge for the model.

A finer-grained breakdown by sentiment and script (Figure 7) reveals that negative comments written in Latin script suffer the highest error rate: 41.7% were misclassified, far exceeding all other categories. In contrast, negative Arabic and mixed comments had error rates of 12.0% and 19.0%, respectively, while positive comments showed more balanced performance across scripts (15.3% for Arabic, 12.2% for Latin, 15.9% for mixed). Given this pronounced vulnerability, we focused our qualitative investigation on the 41.7% of falsely classified negative Latin instances.

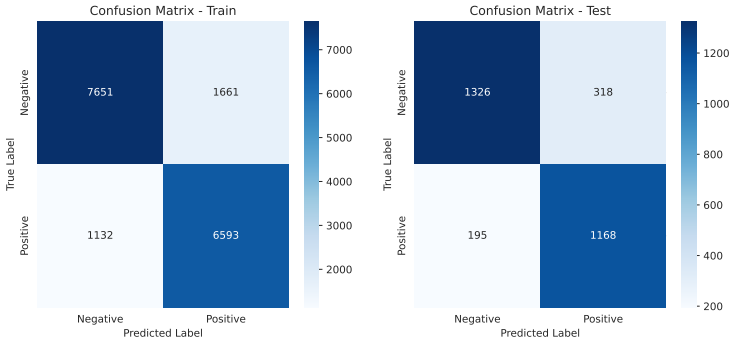


Figure 5. Confusion matrices for the TunDC-mixed model on training and test splits of the TunDC dataset

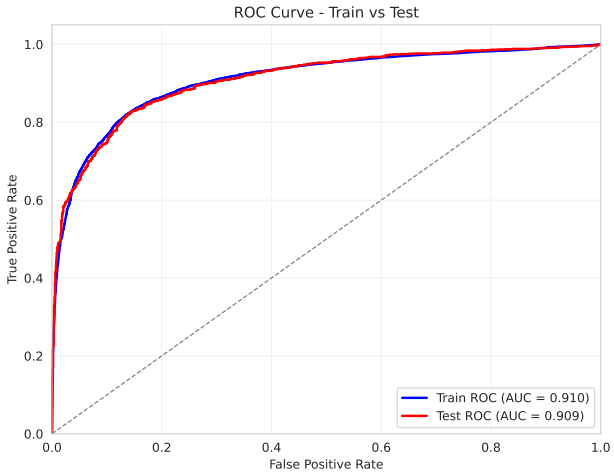


Figure 6. ROC curves for the TunDC-mixed model on training and test sets

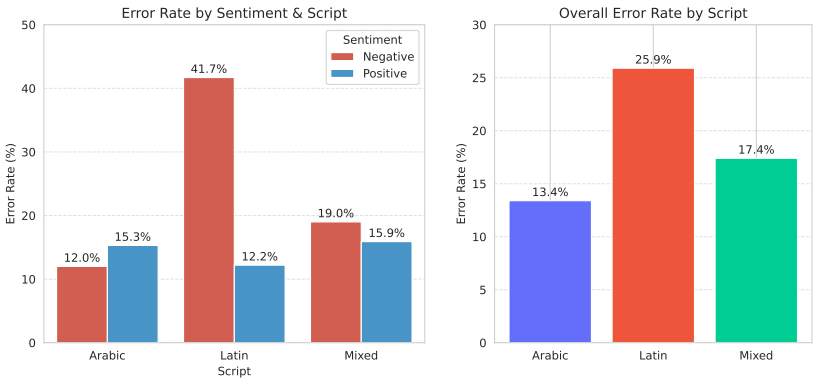


Figure 7. Misclassification rates by sentiment label and writing system on the TunDC test set, highlighting higher error rates for negative comments written in Latin script

We manually examined a random sample of 30 such cases and identified four recurring linguistic phenomena that likely contribute to misclassification:

i) Code-switching: frequent mixing of French, Arabic, and TD expressions disrupts semantic coherence. Examples include:
   - *"khit 3lik tabba3 fi mineur ! dégoûtant .. JE ME DESABONNE !"*

ii) Non-standard orthography: highly phonetic or idiosyncratic spelling obscures lexical identity:
   - *"Santou3et'ha seb9a sh3arha net7adeha tched ch3arha lfou9 hhhhhhh"*,
   - *"wallahimseknat ferhanin itkolech3lehom min3oumitiya7"*.

iii) Sarcasm and irony: negative sentiment is often conveyed indirectly or through mockery:
   - *"Win l9awhom hal ra9asa"*,
   - *"Melkef we lebes balouta ya 7san"*.

To assess whether vocabulary coverage was a factor, we computed out-of-vocabulary (OOV) rates using the tokenizer's [UNK] token frequency. Surprisingly, Latin-script comments exhibited a lower OOV rate (0.26%) compared to Arabic-script comments (0.38%). However, manual inspection of actual tokenization outputs from the test set revealed a more insidious issue: Latin-script inputs undergo excessive subword fragmentation, which degrades semantic coherence despite low OOV counts. Consider the following real token sequences produced by the asafaya/bert-base-arabic tokenizer.

Latin-script examples:

- `['kh', '##it', '3', '##li', '##k', 'tab', '##ba', '##3', 'fi', 'min', '##e', '##ur', '!', 'de', '##go', '##ut', '##ant', '.', '.', 'j', '##e', 'me', 'des', '##ab', '##on', '##ne', '!']`
- `['me', '##l', '##ke', '##f', 'we', 'leb', '##es', 'b', '##al', '##out', '##a', 'y', '##a', '7', '##sa', '##n']`
- `['n', '##a', '##9', '##es', 'ch', '##way', '##a', 'men', 'ma', '##qu', '##ill', '##age', 's', '##i', 'al', '##a', 'h', '##h', '##h']`

While Arabic-script text is largely tokenized at the word or morpheme level (e.g., 'اسمع', 'اخبار'), Latin-script utterances are broken into highly granular, often phoneme-like units (e.g., `'n', '##a', '##9', '##es'` for "na9es") meaning "reduce" in TD. This fragmentation disrupts the model's ability to recognize lexical and pragmatic cues—especially critical for detecting sarcasm, implicit negativity, or code-switched expressions. Thus, the core issue is not lexical coverage but semantic erosion during tokenization, which disproportionately affects Latin-transcribed dialectal text.

We also analyzed the distribution of word counts for correctly versus incorrectly classified samples. Both the raw and log-scaled histograms (Figure 8) show nearly identical distributions, indicating that text length is not a distinguishing factor in model errors. Both linear as in Figures 8(a) and log-scaled as in Figures 8(b) views show overlapping distributions, suggesting length is not a primary error driver. This further supports the hypothesis that misclassifications stem from linguistic complexity and tokenization artifacts rather than surface-level features. These findings underscore a key limitation of current AraBERT models when applied to Latin-transcribed dialectal text: even with adequate lexical coverage, poor tokenization granularity can undermine performance, particularly for sentiment expressions that rely on pragmatic or contextual cues.

4.4. Discussions and insights

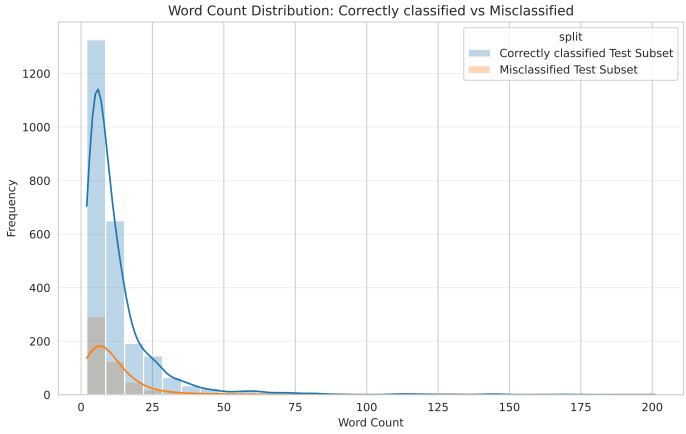
The evaluation results provide valuable insights into the strengths and limitations of bert-base-Arabic-TunDC-mixed in processing TD sentiment analysis. One of the main challenges was writing system variability, where comments often contained a mix of Arabic and Latin scripts, sometimes within the same sentence. While the transformer-based model was capable of handling such variability better than traditional approaches, certain edge cases remained difficult to classify.

Another challenge was dialectal diversity, as the TD exhibits significant variations based on age, region, and social context. These variations resulted in inconsistent spellings, non-standard word usage, and informal grammatical structures. While the model successfully captured many of these patterns, further improvements could be achieved through domain-specific pretraining on larger dialectal corpora.

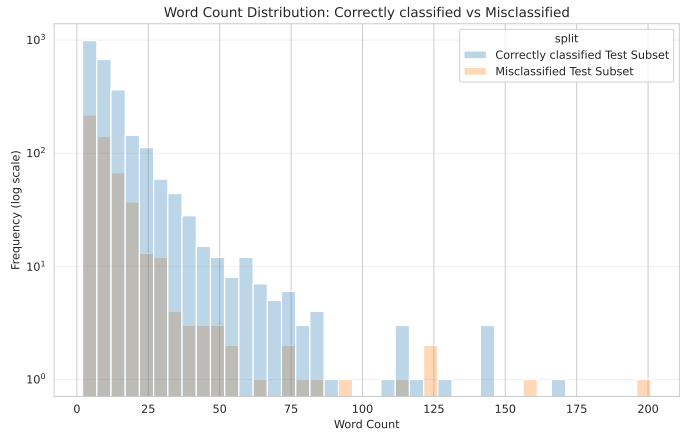
Our error analysis revealed that misclassifications were not uniformly distributed but concentrated in specific linguistic and orthographic contexts. The highest error rate (41.7%) occurred among negative

comments written in Latin script, a pattern driven not by vocabulary gaps (OOV rate: 0.26%) but by severe tokenization fragmentation that breaks words into semantically opaque subunits (e.g., "earing" → `['b', '##al', '##out', '##a']`). This fragmentation impedes the model's ability to interpret pragmatic phenomena such as sarcasm, code-switching, and implicit aggression, which are prevalent in negative Latin-script utterances. In contrast, Arabic-script text, despite a slightly higher OOV rate (0.38%), benefits from more holistic tokenization that preserves morphological integrity. Furthermore, error rates were independent of text length, as confirmed by overlapping word-count distributions between correct and incorrect predictions. These findings highlight that the primary bottleneck lies not in model capacity or data quantity, but in the mismatch between the tokenizer's design, optimized for MSA and Arabic script, and the realities of informal Latin-transcribed TD.

Additionally, short and ambiguous comments were difficult to classify, particularly those that lacked clear sentiment indicators. The model benefited from context-aware embeddings, which helped in many cases; however, further refinements, such as incorporating self-supervised learning techniques, could improve sentiment classification in highly ambiguous texts. The strong performance of bert-base-arabic-TunDC-mixed underscores the importance of fine-tuned pre-trained models for low-resource Arabic dialect NLP tasks. However, additional research is needed to enhance its ability to generalize across broader TD variations, particularly in more conversational or informal settings. The results of these experiments confirm that bert-base-arabic-TunDC-mixed performs effectively on TD sentiment classification, achieving an accuracy of 84% and an F1-score of 0.84. The model demonstrated robust performance in handling dialectal variations, informal spellings, and mixed-script text, benefiting from contextual embeddings learned during fine-tuning.



(a)



(b)

Figure 8. Distribution of word counts for (a) word count (linear scale) and (b) word count (log scale)

5. AI ETHICS AND BIAS IN TUNISIAN DIALECT NLP

The development of NLP resources and models for low-resource dialects like TD is not only a technical challenge but also an ethical one. As the TunDC corpus contributes to advancing sentiment analysis in TD, it is crucial to address potential ethical concerns related to data collection, annotation, and model deployment. Particularly concerning bias and fairness.

5.1. Annotation bias and sentiment skew

A primary ethical concern in dialectal NLP is the potential for annotation bias. The process of manually labeling data, especially for subjective tasks like sentiment analysis, is inherently prone to the annotator's personal, cultural, or political viewpoints. In the context of the TD, this is compounded by the lack of orthographic standardization and the high degree of code-mixing. For instance, an annotator's decision to exclude or normalize code-mixed content (as seen in some prior work [50] can introduce a systemic bias against certain linguistic expressions, which may correlate with specific social groups or topics.

Furthermore, sentiment skew or class imbalance is a common issue in real-world social media data, where one sentiment class (e.g., negative) may dominate. While techniques like stratified sampling can mitigate this during model training, the underlying data distribution may still reflect societal biases. Models trained on skewed data can perpetuate and amplify these biases, leading to unfair or inaccurate predictions for underrepresented groups or sentiments. Recent research emphasizes the need for culturally sensitive annotation criteria to ensure that the resulting models are fair and reliable [51].

5.2. Fairness and inclusivity

The concept of fairness in TA NLP is multifaceted, primarily revolving around linguistic and social inclusivity.

- Linguistic fairness: this relates to the model's performance across different linguistic variations. Given the multi-script nature of TD (Arabic script, Arabizi, and mixed), a model that performs significantly better on one script over another is linguistically unfair. The focus on multi-script robustness in TunDC is a step toward addressing this.
- Social fairness: this addresses the risk of models perpetuating social biases (e.g., gender, regional, or political) present in the training data. For example, if the corpus over-represents political discussions from a specific region, the resulting model may exhibit a political or regional bias when deployed. The ethical deployment of AI in the Arabic-speaking world requires careful consideration of these cultural and social nuances to prevent the marginalization of certain communities [52].

By acknowledging these ethical challenges, researchers can move towards developing more robust, inclusive, and responsible NLP systems for the TD. This includes transparent reporting of annotation guidelines, inter-annotator agreement, and detailed analysis of model performance across different demographic and linguistic subgroups.

6. CONCLUSION AND FUTURE WORK

In this research, we introduced TunDC, a novel and publicly available dataset comprising over 20,000 manually annotated comments, specifically curated to address resource scarcity for the TD. The experimental results validate the utility of this resource, as a fine-tuned BERT-based model achieved a promising accuracy of 84%, highlighting the effectiveness of transformer architectures in capturing the dialect's complex linguistic nuances. By contributing this open benchmark to the research community, this work aim to enable further advancements in sentiment analysis and other NLP applications for low-resource Arabic dialects. Future work will prioritize the creation of a more representative and balanced TD dataset that explicitly accounts for script usage, regional variation, and pragmatic phenomena. This includes collecting more Latin-script negative examples—particularly those exhibiting sarcasm, code-switching, and implicit sentiment—and ensuring proportional representation across writing systems. Additionally, incorporating metadata such as region, age group, and platform context could enable more nuanced modeling. Finally, developing a parallel corpus with standardized orthographic normalization (e.g., mapping Latin-script Tunisian to a canonical form) would help mitigate tokenization-induced semantic loss while preserving linguistic authenticity.

ACKNOWLEDGEMENT

The authors extend their appreciation to Umm Al-Qura University, Saudi Arabia for funding this research work through grant number: 26UQU4350491GSSR01.

FUNDING INFORMATION

This research work was funded by Umm Al-Qura University, Saudi Arabia, under grant number: 26UQU4350491GSSR01.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ahmed Khalil Boulahia	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Mourad Mars	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject Administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding Acquisition
Fo	: **Fo**rmal Analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Derived data supporting the findings of this study are available from the corresponding author, [MM], upon request. However, the trained classification models are available on Hugging Face.

REFERENCES

[1] M. Mars, "From word embeddings to pre-trained language models: a state-of-the-art walkthrough," *Applied Sciences*, vol. 12, no. 17, 2022, doi: 10.3390/app12178805.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, 2019, pp. 4171–4186.

[3] B. Alharbi *et al.*, "ASAD: a Twitter-based benchmark Arabic sentiment analysis dataset," 2020, arXiv:2011.00578.

[4] A. Messaoudi, H. Haddad, M. B. HajHmida, C. Fourati, and A. B. Hamida, "Learning word representations for Tunisian sentiment analysis," *Pattern Recognition and Artificial Intelligence*, vol. 1322, pp. 329–340, 2020, doi: 10.1007/978-3-030-71804-6_24.

[5] A. Abdelali, H. Mubarak, Y. Samih, S. Hassan, and K. Darwish, "Arabic dialect identification in the wild," 2020, arXiv:2005.06557.

[6] M. Nabil, M. Aly, and A. F. Atiya, "ASTD: Arabic sentiment tweets dataset," *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2515–2519, doi: 10.18653/v1/d15-1299.

[7] S. Shon, A. Ali, Y. Samih, H. Mubarak, and J. Glass, "Adi17: A fine-grained Arabic dialect identification dataset," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020, pp. 8244–8248, doi: 10.1109/ICASSP40776.2020.9052982.

[8] E. Boujou, H. Chataoui, A. E. Mekki, S. Benjelloun, I. Chairi, and I. Berrada, "An open access NLP dataset for Arabic dialects: data collection, labeling, and model construction," 2021, arXiv:2102.11000.

[9] E. Fsih, S. Kchaou, R. Boujelbane, and L. H. Belguith, "Benchmarking transfer learning approaches for sentiment analysis of Arabic dialect," *WANLP 2022 - 7th Arabic Natural Language Processing - Proceedings of the Workshop*, 2022, pp. 431–435, doi: 10.18653/v1/2022.wanlp-1.44.

[10] Y. Matrane, F. Benabbou, and N. Sael, "A systematic literature review of Arabic dialect sentiment analysis," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 6, 2023, doi: 10.1016/j.jksuci.2023.101570.

[11] F. Husain, H. Al-Ostad, and H. Omar, "A weak supervised transfer learning approach for sentiment analysis to the Kuwaiti dialec," *WANLP 2022 - 7th Arabic Natural Language Processing - Proceedings of the Workshop*, 2022, pp. 161–173, doi: 10.18653/v1/2022.wanlp-1.15.

[12] M. Mhamed, R. Sutcliffe, and J. Feng, "Benchmark Arabic news posts and analyzes Arabic sentiment through RMuBERT and SSL with AMCFFL technique," *Egyptian Informatics Journal*, vol. 29, 2025, doi: 10.1016/j.eij.2024.100601.

[13] A. Vavre, A. Gupta, and S. Sarawagi, "Adapting multilingual models for code-mixed translation," *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 7162–7170, doi: 10.18653/v1/2022.findings-emnlp.485.
[14] Y. Noureldien, A. Mohamed, and F. Attallah, "Zero-shot and fine-tuned evaluation of generative LLMs for Arabic word sense disambiguation," in *Proceedings of The Third Arabic Natural Language Processing Conference*, 2025, pp. 298–305, doi: 10.18653/v1/2025.arabicnlp-main.24.
[15] H. Al-Owais and A. Elnagar, "Arabic dialect detection using large language models: a comparative analysis," *2025 International Conference on New Trends in Computing Sciences (ICTCS)*, 2025, pp. 454–459, doi: 10.1109/ICTCS65341.2025.10989347.
[16] A. Dahou *et al.*, "A survey on dialect Arabic processing and analysis: recent advances and future trends," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 8, 2025, doi: 10.1145/3747290.
[17] S. B. Hassine, A. Arrak, M. Addhoum, and S. R. Wilson, "TounsiBench: benchmarking large language models for Tunisian Arabic," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 34615–34630, doi: 10.18653/v1/2025.emnlp-main.1756.
[18] E. Gugliotta and M. Dinarelli, "TArC: Tunisian Arabish corpus first complete release," *2022 Language Resources and Evaluation Conference (LREC)*, 2022, pp. 1125–1136.
[19] H. Naouara, J.-P. Lorré, and J. Louradour, "LinTO audio and textual datasets to train and evaluate automatic speech recognition in Tunisian Arabic dialect," 2025, arXiv:2504.02604.
[20] S. O. Alhumoud, M. I. Altuwaijri, T. M. Albuhairi, and W. M. Alohaideb, "Survey on Arabic sentiment analysis in Twitter," *International Journal of Social, Behavioral, Educational, Economic and Management Engineering*, vol. 9, no. 1, pp. 364–368, 2015.
[21] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.
[22] Y. Bao, C. Quan, L. Wang, and F. Ren, "The role of pre-processing in Twitter sentiment analysis," in *Intelligent Computing Methodologies (ICIC 2014)*, 2014, pp. 615–624, doi: 10.1007/978-3-319-09339-0_62.
[23] K. Darwish *et al.*, "A panoramic survey of natural language processing in the Arab world," *Communications of the ACM*, vol. 64, no. 4, pp. 72–81, 2021, doi: 10.1145/3447735.
[24] A. Assiri, A. Emam, and H. Aldossari, "Arabic sentiment analysis: a survey," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 12, 2015, doi: 10.14569/ijacsa.2015.061211.
[25] M. Jarrar, N. Habash, F. Alrimawi, D. Akra, and N. Zalmout, "Curras: an annotated corpus for the Palestinian Arabic dialect," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 745–775, 2017, doi: 10.1007/s10579-016-9370-7.
[26] J. Pearl, *Probabilistic reasoning in intelligent systems, networks of plausible inference*. New York, United States: Morgan Kaufmann, 1988.
[27] N. AlHazzani *et al.*, "Sa'7r: a Saudi dialect irony dataset," *5th Workshop Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, OSACT 2022 - Proceedings at Language Resources and Evaluation Conference, LREC 2022*, 2022, pp. 60–70.
[28] S. Khalifa, N. Habash, D. Abdulrahim, and S. Hassan, "A large scale corpus of Gulf Arabic," *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016, pp. 4282–4289.
[29] S. Harrat, K. Meftouh, K. Smaïli, and K. S. M. Arabic, "Maghrebi Arabic dialect processing: an overview," *Journal of the International Science and General Applications*, vol. 1, no. 1, 2018.
[30] M. A. Mageed, C. Zhang, A. R. Elmadany, H. Bouamor, and N. Habash, "NADI 2021: the second nuanced Arabic dialect identification shared task," *WANLP 2021 - 6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop*, 2021, pp. 244–259.
[31] M. A. Mageed, A. R. Elmadany, C. Zhang, E. M. B. Nagoudi, H. Bouamor, and N. Habash, "NADI 2023: the fourth nuanced Arabic dialect identification shared task," *ArabicNLP 2023 - 1st Arabic Natural Language Processing Conference, Proceedings*, 2023, pp. 600-613.
[32] K. Sayadi, M. Liwicki, R. Ingold, and M. Bui, "Tunisian dialect and modern standard Arabic dataset for sentiment analysis: Tunisian election context," *Proceedings of the 17th International Conference on Intelligent Text Processing and Arabic Computational Linguistics*, 2016.
[33] N. Karmani, "Tunisian Arabic customer's reviews processing and analysis for an-internet supervision system," Ph.D. dissertation, Department of Computer System Engineering, Sfax University, National Engineering School of Sfax, 2017.
[34] S. Medhaffar, F. Bougares, Y. Estève, and L. H. Belguith, "Sentiment analysis of Tunisian dialects: linguistic resources and experiments," *Proceedings of the Third Arabic Natural Language Processing Workshop*, 2017, pp. 55–61, doi: 10.18653/v1/W17-1307.
[35] C. Fourati, A. Messaoudi, and H. Haddad, "TUNIZI: a Tunisian Arabizi sentiment analysis dataset," 2020, arXiv:2004.14303.
[36] A. Masmoudi, J. Hamdi, and L. H. Belguith, "Deep learning for sentiment analysis of Tunisian dialect," *Computacion y Sistemas*, vol. 25, no. 1, pp. 129–148, 2021, doi: 10.13053/CYS-25-1-3472.
[37] H. Mulki, H. Haddad, C. B. Ali, and I. Babaoglu, "Tunisian dialect sentiment analysis: a natural language processing-based approach," *Computacion y Sistemas*, vol. 22, no. 4, pp. 1223–1232, 2018, doi: 10.13053/CyS-22-4-3009.
[38] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, vol. 2, pp. 427–431, 2017, doi: 10.18653/v1/e17-2068.
[39] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.zip: compressing text classification models," 2016, arXiv:1612.03651.
[40] K. Lee *et al.*, "Deduplicate-text-datasets," *GitHub*. [Online]. Available: https://github.com/google-research/deduplicate-text-datasets
[41] K. Lee *et al.*, "Deduplicating training data makes language models better," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2022, pp. 8424–8445, doi: 10.18653/v1/2022.acl-long.577.
[42] M. Allamanis, "The adverse effects of code duplication in machine learning models of code," *Onward! 2019 - Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software, co-located with SPLASH 2019*, 2019, pp. 143–153, doi: 10.1145/3359591.3359735.

[43] J. Bandy and N. Vincent, "Addressing 'documentation debt' in machine learning: a retrospective datasheet for BookCorpus," *Advances in Neural Information Processing Systems*, 2021.
[44] A. Vicenzi, T. Gringauz, and Harry, "Emojis," *GitHub*. [Online]. Available: https://github.com/alexandrevicenzi/emojis
[45] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, pp. 378–382, 1971, doi: 10.1037/h0031619.
[46] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[47] A. K. Boulahia, "Bert-base-arabic-tundc-mixed," *Hugging Face*. [Online]. Available: https://huggingface.co/AhmedBou/bert-base-arabic-TunDC-mixed
[48] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media," *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings*, 2020, pp. 2054–2059, doi: 10.18653/v1/2020.semeval-1.271.
[49] A. Safaya, "Bert-base-arabic," *Hugging Face*. [Online]. Available: https://huggingface.co/asafaya/bert-base-arabic
[50] C. Fourati, R. Hammami, C. Latiri, and H. Haddad, "PoliTun: Tunisian political dataset for detecting public opinions and categories orientation," *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, 2024, pp. 178–185.
[51] H. Rhel and D. Roussinov, "Large language models and Arabic content: a review," in *Selected Papers from the International Conference on Artificial Intelligence (FICAILY 2025)*, 2026, pp. 402–419, doi: 10.1007/978-3-032-00232-7_26.
[52] E. Ferrara, "Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, Dec. 2023, doi: 10.3390/sci6010003.

BIOGRAPHIES OF AUTHORS

Ahmed Khalil Boulahia holds a master's degree in artificial intelligence, systems, and data from the Université Paris Dauphine – Tunis (2020). He is currently a machine learning engineer at Mantu, based in Tunis, Tunisia, where he works on developing data-driven solutions using machine learning and natural language processing techniques. His areas of interest include deep learning, large language models, generative AI, and the integration of AI systems into real-world applications. He is currently engaged in independent research and personal AI projects aimed at exploring innovative applications of artificial intelligence to solve practical challenges and advance his expertise in the field. He can be contacted at email: ahmed.boulahia@dauphine.eu.

Mourad Mars received a Master of Science (M.Sc.) degree and a Doctor of Philosophy (Ph.D.) degree in artificial intelligence (AI) and natural language processing (NLP) from Grenoble Alpes University, Grenoble, France, in 2006 and 2012, respectively. His major field of study is AI and NLP. He has held various academic, managerial, and technical roles, including positions as a lecturer, researcher, and head of department for several years. In addition to his academic background, he holds a project management professional (PMP) certification, reflecting his expertise in managing complex research and development projects. He currently serves as an assistant professor in the Department of Computer Science and Artificial Intelligence at the College of Computing, Umm Al-Qura University (UQU), Mecca, Saudi Arabia. In addition to his teaching and research duties, he has supervised several master's and Ph.D. students in AI-related research. He has published numerous scientific papers in peer-reviewed journals and international conferences and actively serves as a reviewer for academic journals and leading AI/NLP conferences. His research interests span generative AI, large language models, retrieval-augmented generation (RAG), AI agents, automatic speech recognition, multilingual NLP, machine translation, and AI applications in education and industry. He serves as the principal investigator (PI), SPI, and Co-PI on several funded, AI-driven research projects in collaboration with industry and government entities, including the Saudi Ministry of Culture and the Research, Development, and Innovation Authority (RDIA) of Saudi Arabia, among others. He can be contacted at email: msmars@uqu.edu.sa.